典籍事件触发动词识别研究:基于《左传》的文本实验*

何琳[1,2] 马晓雯[1,3] 喻雪寒[1,2] 艾毓茜[1,2] 李章超[1,2] 高丹[1,2]

[1] 南京农业大学信息管理学院 南京 210095 [2]南京农业大学人文与社会计算研究中心 南京 210095

[3] 南京医科大学图书馆 南京 210029

摘 要:[目的/意义]事件自动识别抽取是当前典籍主题挖掘研究中一个新的重要课题,其中事件触发词的识别是一项基础的工作,本研究旨在探索古代典籍中事件触发词自动识别和分类的通用方法。[方法/过程]首先运用 LDA 模型对动词进行主题聚类,归纳典籍事件触发动词的分类体系;并依据聚类结果与分类体系,初步构建触发动词的种子词集。在此基础上,通过语义相似度计算,对种子词集进行扩展,构建典籍事件触发词语义数据集。在实验阶段,以先秦时期的重要典籍《左传》为例,对分类体系构建和种子词集扩展的方法进行验证。[结果/结论]结果表明,本文所提出的典籍事件触发词识别方法可行有效,据此构建的事件触发词集具有较高可信度,未来可进一步扩大实验的样本数量及范围。

关键词:触发词识别 主题聚类 词集扩展 类别体系构建 典籍文本

分类号:G250

DOI:10.13266/j.issn.0252-3116.2022.05.014

1 引言

随着古籍文本数字化资源的快速增长以及文本挖掘技术和人文计算工具的不断进步,古文信息处理的研究也日益朝着智能化与深语义化的方向发展。如何准确、有效地从古文中提取和挖掘出广覆盖、多层次、有价值的知识是古文信息处理研究的重要任务之一[1]。在典籍深层次文本分析与挖掘中,需要将古籍文本中的人名、地名、事件、时间等具体命名实体信息进行提取,进而发现这些不同命名实体之间的语义关系,实现典籍文本的深度标注与知识关联,构建典籍知识库,在此基础上探究各种历史事件在时间和空间上的演变规律。在这一过程中,事件识别与抽取是实现典籍文本细粒度组织的重要手段,对于典籍文本知识库构建质量具有重要价值,其中事件触发词的识别又是决定事件抽取与识别效果的一项基本而关键的工作。有研究表明,超过60%的事件抽取错误是由于触发词识别过程中的错误导致[2]。

触发词是能够表征事件发生的词语,触发词识别的过程本质上就是通过触发词的自动抽取和分类,判定事件类别的过程。目前在面向一些特定领域的现代文本研究中,事件触发词的识别取得了较好的效果,比如在金融领域[3]、音乐领域[4]以及突发事件[5]的识别和抽取方面)。然而在面对古籍文本时,由于古籍行文结构和句法的特殊性,缺乏通用的触发词抽取规则,因此在基础词典的构建方面有待进一步的探索,也尚未建立起完备的典籍触发词分类体系,在触发词的识别和类别判定上存在一定的难度。

本文结合现代文本触发词识别取得的进展,借助触发词识别在特定领域研究中采用的技术,探索典籍文本中事件触发词的自动识别和事件抽取方法。本文首先依据古文语法特征,运用 LDA 模型对典籍动词进行了主题聚类,归纳了主题类别,构建了典籍触发动词分类体系;随后根据分类体系和聚类的结果构建了触发词种子词集,抽取了词典语义特征及动词上下文特征,进而利用语义相似度计算等文本挖掘技术对种子词集进行了扩展;最后本文对提出的触发词分类体系和数据集构建的方法均进行了实证研究,并对结果进行了人工校验与一致性检验。结果表明,本文所提出的典籍事件触发词数据集构建方法具有较高的可信度。

* 本文系国家社会科学基金项目"基于典籍的中华传统文化知识表达体系自动构建方法"(项目编号:18BTQ063)研究成果之一。

作者简介:何琳,教授,博士生导师,E-mail:helin@njau.edu.cn;马晓雯,硕士研究生;喻雪寒,博士研究生;艾毓茜,硕士研究生;李章超,博士研究生;高丹,博士研究生。

收稿日期:2021-08-01 修回日期:2021-11-21 本文起止页码:133-141 本文责任编辑:杜杏叶

2 相关研究综述

2.1 触发词识别抽取的相关研究

事件触发词指能够表征事件发生的词，是决定事件类型最重要的特征词。事件触发词抽取包括触发词检测与分类，首先判定当前句子是否存在事件触发词以实现事件检测，然后通过识别事件触发词类型判断事件类型。在信息抽取领域，事件触发词的识别方法主要有三种：基于统计的方法、基于规则的方法和机器学习方法。

基于统计的方法是人工统计出句子或文本中的所有触发词，建立一个触发词词典，通过词典来判断其他词语是否为触发词[6]。这种方法简单且技术要求不高，但它是一种典型的经验性方法，要求训练语料规模足够大且经典，因此受到语料的限制并不能保证统计和测试结果的正确性，统计过程也费时费力[7]。

基于规则的方法则是事先定义一些规则寻找触发词。付剑锋[8]研究得出触发词一般是动词或名词的结论，以此规则过滤掉其它词性的词语。这一方法能有效地提高触发词的识别效率，但依赖初始规则的定义，并且难以涵盖所有特征，可能过滤掉一些本身可以充当触发词的词，导致识别效果较低[9]。规则的定义过程会耗费大量人力，并且往往针对特定数据集定义，特定数据集的规则或模式难以应用到其他的数据集，泛化性能较差[10]。

基于机器学习的方法是基于训练集进行自动学习[11]，是目前研究较多使用也最为广泛的一种方式，它主要利用特征集训练触发词识别分类器，把触发词的识别问题转化为了分类问题[12]。陈箫箫[13]等运用条件随机场模型实现了事件抽取中的序列标记，利用LDA主题模型建立了事件抽取和分类系统；景悦诚[14]实现了从数据抓取及预处理到人工标注再到机器学习，最终实现事件挖掘的完整流程，并对新浪微博进行了文本挖掘。机器学习的方法也存在着不足，一方面需要足够量的特征集训练分类器，以保证识别结果的精确率，这要求训练语料和测试语料都达到一定的规模；另一方面，模型特征的有效性决定着系统的性能，大多数研究也致力于为事件抽取构造丰富有效的特征集，从而提升事件抽取方法的性能[15]。

2.2 古文信息处理的相关研究

古文信息处理就是运用自然语言处理技术，对数字化的古籍文本进行分词与词性标注、命名实体识别以及古籍语料库构建等工作，在此技术上对古籍文本做关系抽取、深度挖掘和可视化展示[16]，实现典籍数据的有效组织和利用。

在古文分词方面，分为基于规则的人工方法与基于统计的机器学习方法。基于规则的方法主要是针对一些语法、句式等都有相同规律或特征的结构化文本，如邱冰[17]等将《汉语大词典》作为通用分词词典，通过逆向匹配进行古文分词；徐润华[18]等利用古籍注疏文献的词汇语义知识，通过文献和注疏自动对齐的方式对《左传》进行分词。基于机器学习的方法从语料中提取文本特征，在标注语料上训练模型之后对对未标注语料实施自动分词，大大提高了效率。王嘉灵[19]等采用最大匹配法结合《汉书》注疏表，并利用CRF模型对《汉书》做了全面系统的分词研究；F. Chen等[20]构建了交互式古文在线自动分词平台，根据用户反馈实时优化分词性能。

古文的词性标注主要是利用机器学习方法进行，常用的模型有CRF、HMM序列标注模型、Bi-LSTM深度学习模型等，目前已经建立了许多标注语料库。黄建年等[21]利用计算机技术构建了农业古籍断句标点、分词标引的原型系统。陈小荷[22]建立了古籍自动分析工具和先秦古汉语标注语料库。台湾研究院[23]建立的汉籍电子文献“瀚典全文检索系统”，其中包含汉籍全文资料库、古汉语语料库、近代史全文资料库等18个子库。

古文的事件抽取研究方面，利用机器学习和深度学习方法对事件命名实体识别的研究取得了一些成果。王东波[24]等对先秦典籍中历史事件基本实体做了内部的数量统计和外部特征分析之后，构建了特征模板。但针对古文中事件抽取前的触发词自动识别阶段的研究相对较少，大多数研究重点在于对事件要素和事件联系的挖掘，如刘忠宝等[25]建立了一个可视化图谱系统，对《史记》中的历史事件及其组成要素进行了自动抽取，并将不同事件的关系在知识图谱中展现出来。

综上所述，在古文的自动分词及词性标注方面已经有了丰富的研究成果。而事件实体的识别和事件关系抽取等深层次文本挖掘处理，研究还较为薄弱，限制了古代文本资源加工与整理的层次。究其原因，一方面是缺乏古代汉语文本深度标注语料库，另一方面是缺乏触发动词识别等的基础性研究。

现代汉语触发词识别和事件关系抽取中取得了较好的成果，这些研究为面向典籍的触发动词识别提供了方法论支撑。由于古代汉语语法的特征，给古代汉

语文本处理带了一定的壁垒，令人欣喜的是当前古代汉语在分词、命名实体识别等领域取得了较好的成果，这些都为面向古文本的触发动词识别奠定了良好的基础。本文正是在此基础上，对典籍文本中触发动词的分布规律与特征进行综合性研究，进而利用自然语言处理技术探索触发动词分类体系构建的可行性，并以小规模典籍为例，初步验证了触发动词数据集在典籍事件抽取中的效果，为典籍文本深度挖掘提供借鉴与参考。

3 研究方法

3.1 总体研究框架

本文总体研究框架如图 1 所示，分为典籍文本预处理、触发动词分类体系构建、触发动词集扩展及结果验证四个部分。

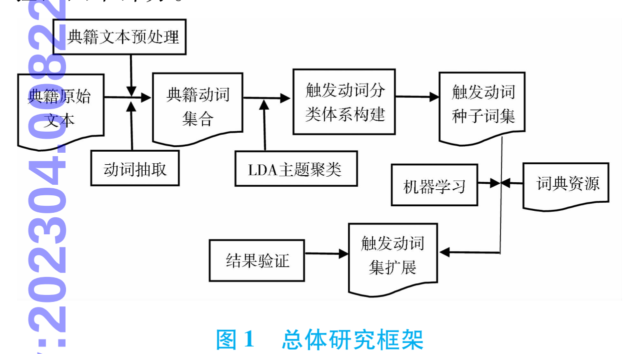


图 1 总体研究框架

首先通过语料筛选、去停用词、动词抽取等步骤对原始典籍语料做预处理；其次运用 LDA 聚类模型对抽取出的典籍动词进行词义聚类，结合了数量统计与定性分析的方法，构建触发词分类体系；最后将聚类结果中的特征词集合起来形成触发词种子词集，通过机器学习与词典相结合的方式对词集进行同义词扩展，形成完整的典籍事件触发动词集。在实验阶段，选取《左传》文本对上述方法进行了验证，将触发动词词表与事件句数据集对应，在此基础上对事件句文本做结构化标注，通过误差验证与一致性检验验证触发动词数据集的有效性。

3.2 典籍事件触发词分类体系的构建依据

由于客观事物本身具有多种属性和多维度联系，任何一种分类体系都存在自身的分类依据，因此典籍蕴含的事件也存在多种划分维度。人文学科对于典籍文本的研究中，形成了关于社会事件主题的研究成果。对于典籍词汇的研究，语言学领域建立了标准词汇场研究词汇对社会现象及事件特征的揭示。《ACE 中文事件指南》[26]把“事件”的特点定义为包含参与者、与人物特定活动相关、存在一定的状态变化。结合上述研究，本文将关于典籍主题分析和词汇类别研究的相关方法，作为构建触发词分类体系所参考的理论依据。

3.2.1 典籍的主题分布规律

从语言的外部环境来看，词汇是社会生活中最生动、最客观的反映。因此，一定历史时期的政治、经济、军事、外交、风俗习惯、规章制度等的变化，都会在典籍文本的词汇中显性或隐性地反映出来[27]。基于这种认识，发现典籍中词语的内部联系，探索建立有效的词义系统和分类体系，将这些词汇置于特定的社会文化背景和语境中进行分析考察，能够有规律地揭示典籍文本所反映的社会主题。

3.2.2 典籍的动词语义场

从语言学的角度来看，想要快速准确地了解一个时代的社会发展，从“标志性词汇场”和“基本词汇”入手，是一种切实有效的途径[28]。一个孤立的词反映社会现象和特征的能力是有限的，但当多个语义相关的词被聚合时，就可以鲜明有效地反映具有相似特征的社会现象和事件类别。根据词汇的相同特征和相互关系，将词汇分为不同的类别，实现词义的聚合。

因此，本文选取代表性的典籍，通过聚类等典籍内容分析方法实现对面向典籍的触发动词分类体系构建。根据对动词的统计结果，建立一种自下而上类别生成方式，构建触发动词分类体系。

3.3 典籍文本预处理方法

预处理阶段，针对古籍的特点去除停用词，并提取出语料中的动词作为后续主题聚类和词集构建的基础。

3.3.1 语料筛选

本文的研究是针对典籍中所涉及的事件触发动词进行分类识别，因此重点关注原始文本中描述具体事件或阐述客观事实的内容。由于典籍文本内容的复杂性，如对话类、时间类、引用类的文本暂时不作为备选语料。

3.3.2 去停用词

典籍文本中包含许多没有实际含义的介词、连词、量词及古文所特有的词缀等，影响主题判断的准确度和文本内容分析的质量。采用了汉典古籍停用词表，共包含 187 个停用词，如“以、诸、之、曰”等。根据这一停用词典对典籍语料进行进一步的处理。

3.3.3 动词抽取

触发词分类体系主要是针对文本中的动词，本文结合词性标注的结果，利用正则表达式获取原始语料

中动词。在抽取出的动词中存在同字异形的情况,如“为”与“爲”,将这些重复的字词进行综合去重,并将能愿动词、关系动词、存现动词等一些情况进行筛除,对结果进行词频统计和整理,获得典籍文本动词的词频分布表。

3.4 触发动词分类体系构建方法

本文利用 LDA 模型对抽取出的动词进行了主题聚类,并对聚类结果进行横向和纵向的对比以及对内容定性分析,确定主题数量,并对主题内容做有效归纳,赋予主题词。

3.4.1 LDA 主题聚类框架

LDA 主题模型是一个基于贝叶斯统计的模型,它采用无监督学习以及词袋(bag of words)的方法对语料库中隐含的主题建模,将每篇文档看作是一条词频向量,将文本信息数字化之后通过计算机进行建模和计算。本文采用 LDA 主题模型进行动词主题聚类的原理如图 2 所示:

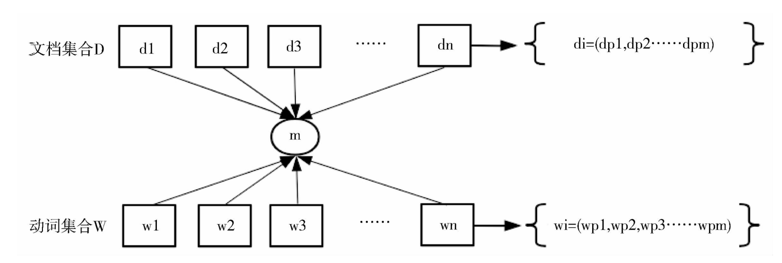


图 2 动词主题聚类 LDA 模型原理

LDA 模型的算法输入是文档的集合 D = {d1, d2, d3, . . . , dn} 和主题类别数量 m,通过计算每一篇文档 di 在所有 Topic 上的一个概率值 p,得到概率的集合 di = (dp1, dp2, . . . , dpm);同样的,对于文档中所有动词的集合 W 也会求出它对应每个 Topic 的概率,wi = (wp1, wp2, wp3, . . . , wpm)。结果得到两个矩阵:文档到 Topic 以及动词到 Topic,这样 LDA 算法就将文档和词投射到了一组 Topic 上,通过 Topic 找出文档与动词之间,文档与文档之间,动词与动词之间潜在的关系。聚类后通过统计出各个 Topic 上词的概率分布,那些在该 Topic 上概率高的动词,能较好的描述该 Topic 的意义。

3.4.2 确定聚类主题数量

在 LDA 聚类模型算法中,主题数量的选择对于结果有较大影响。主题数量太少会导致语义上关联不大的内容合并到某一个嵌合主题中;主题数量太多会导致属于同一主题的内容分裂为若干个不相关的单独主题,造成主题冗余或“垃圾”主题的产生[29]。

本文在对典籍文本的主题分类研究成果归纳分析基础上,设定了初步的聚类主题数量,并依次递增主题数目得到多次主题聚类的结果,对结果进行横向和纵向的对比。横向对比主要是针对每次实验结果中,每个主题下概率值较高的特征动词与其他主题特征动词之间的语义差别是否显著。纵向对比主要是针对每次实验结果中,同一主题下的不同特征动词之间的语义相似度是否聚合。经过多次实验找到横向和纵向对比结果的平衡,确定合适的聚类主题数量以尽量减少垃圾或嵌合主题的产生。

3.4.3 归纳典籍动词主题

在经过 LDA 主题聚类之后得到了主题 – 特征动词的概率分布,通过对概率分布的定性分析可以对每个主题下特征动词所表征的内容进行有效归纳,以便于揭示其深层语义内涵。通常情况下,主题聚类的结果对于典籍中描述较多、文本信息较为丰富的事件主题有较好的揭示效果,但不同的典籍中往往还有一些特殊的事件类别,需要根据不同典籍的情况做少量的归纳补充来进一步完善触发动词分类体系的构建结果。

3.5 触发动词集扩展方法

本文通过机器学习和词典释义方法对种子词集进行扩展。基于机器学习的方法结合动词上下文特征,计算出动词与动词之间、不同事件句之间的余弦相似度,对结果进行统计分析得到种子词的近义词。基于词典释义的方法将抽取出的动词与典籍词典进行匹配得到对应的释义,对匹配结果中的通假字等进行筛选得到种子词的近义词。

3.5.1 种子词集构建

在 LDA 动词聚类实验的结果中,已经得到了每个主题特征动词的概率分布,根据概率进行排序之后,设定阈值,选出的概率较高的动词就是与该主题具有强相关性的事件触发动词,将这些词集合起来作为不同主题事件触发词的种子词集。

3.5.2 文本特征抽取

特征抽取主要包括对典籍词典中动词的词义特征抽取以及原始文本上下文特征的抽取(1)词典语义特征抽取。本文根据词性及词语出处等，将词典中非动词词性的词语，不明出处的词条，词类活用进行了人工筛除。将这些数据与典籍中抽取出的动词做匹配，最终得到了待分类的备选动词在词典中的词义特征；(2)上下文特征抽取。典籍中的动词多为单字词且句子简短，上下文特征不足影响动词的词间相似度计算效果。因此需要以句子为单位计算相似度，寻找与主题相关的事件触发词。本研究所提取的上下文特征主要包括词性特征和词共现特征。词共现的特征是在对语言学的研究成果进行了系统总结全面、整理的基础上得出的经验性结论，例如，"以、於、诸、自、与、及"七个介词常常与战争、杀戮类的动词共现等。

词性特征是为了在进行相似度计算时对同一句子中不同词性的词赋予不同的权重。陈宏[30]曾对汉语同义并列结构中复合词的词性、词序进行统计分析。结果表明，动词对于揭示文本语义，标识文本特性具有重要地位，在句法关系中也具有重要的作用，而文本中的虚词，揭示文本语义的作用不明显。因此要对动词赋予较高权重，以保证计算结果的准确和有效。本文通过对不同词性在汉语句子结构中的分布以及语法规则的统计，结合典籍文本的语法特性，将不同词性的词赋予相应权重。

3.5.3 语义相似度计算

利用 BERT 深度学习模型中的 keras-bert 函数将典籍语料向量化。分词后句子的首尾分别以[CLS]和[SEP]标记，其中[CLS]位置对应的输出向量是能代表整句的句向量，而[SEP]是句间的分隔符，其余部分都是单字输出。在此基础上，利用余弦距离作为文本相似度度量的方法计算文本间的相似度。图 3 是利用 Keras-bert 函数进行文本向量化处理的流程图。

4 实证研究

4.1 实验数据

古代汉语标注语料库是开展本文实验研究的重要数据来源，由于古代汉语的特殊性，现阶段开放共享的深度标注数据仍然较为缺乏。围绕先秦时期语料库的建设，南京师范大学[22]等研究机构开展了先秦典籍文本标注研究工作，完成了典籍分词、词性标注等工作，该工作全部由语言学领域专业人员进行人工标注，是进行典籍文本挖掘的重要数据来源。本文选取了先秦

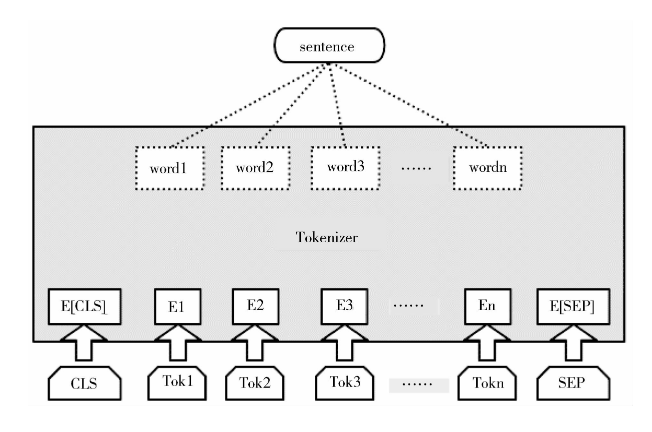


图 3 文本向量化原理

时期的重要典籍、中国较为完备的编年体史书《左传》为实验数据，其跨度了春秋时期 200 余年时间，较好地记录了该时期的社会发展概况。在辅助词典方面，选取了《汉语大词典》《春秋左传词典》《左传详解词典》作为同义词抽取的数据来源，构建了同义词词典。

4.2 触发动词分类体系构建结果

4.2.1 文本预处理

将原始文本预处理后，得到共 17 140 条短句。随后进行动词抽取和去重处理，最终识别出 2 305 个动词，词频分布情况如表 1 所示：

表 1 《左传》动词抽取结果词频统计

词频范围	动词数/个	示例
大于 200	41	使、伐、如、归、杀、入、盟、奔、亡、败
100－200	46	卒、生、谋、帅、救、讨、侵、逆、纳、聘
50－100	72	城、逐、卜、逃、攻、次、御、降、出奔
20－50	175	授、击、封、赂、囚、陈、戮、娶、薨、崩
10－20	197	刑、征、斩、败绩、劫、作乱、缮、即位
3－9	511	莅、弃、筑、贼、送葬、赠
2	273	求成、自杀、莅盟、膳、昏
1	990	—

4.2.2 动词主题聚类

在进行横向对比时发现，聚类主题数量为 8 时，不同主题之间的特征动词语义差别最为显著；在进行纵向对比时发现，聚类主题数量为 8 时，同一主题下的不同特征动词语义相似度较大。因此，将聚类主题数量定为 8 时效果最好，聚类结果见表 2。

结合文学与史学的相关研究成果，并邀请三位领域专家对聚类结果进行评估后，增加了两个主题：社会交换和生活风俗。社会交换包括政治、外交、军事等社会活动中的贿赂行为、馈赠行为、索取行为等。生活风俗主要为祭祀和占卜等社会生活中的重要活动等。

表 2 主题数为 8 时的动词聚类结果

主题 1		主题 2		主题 3		主题 4		主题 5		主题 6		主题 7		主题 8	
贡献度	动词	贡献度	动词	贡献度	动词	贡献度	动词	贡献度	动词	贡献度	动词	贡献度	动词	贡献度	动词
0.057	得	0.083	谓	0.07	告	0.085	从	0.227	使	0.201	为	0.079	伐	0.096	可
0.046	盟	0.074	对	0.042	取	0.058	及	0.079	杀	0.044	奔	0.055	立	0.073	无
0.034	能	0.062	在	0.029	问	0.031	会	0.042	败	0.027	止	0.055	知	0.069	如
0.032	获	0.062	见	0.02	怒	0.027	救	0.031	来	0.025	欲	0.052	执	0.063	死
0.03	卒	0.048	言	0.019	纳	0.027	亡	0.026	复	0.023	战	0.043	事	0.048	归
0.029	行	0.022	舍	0.018	恶	0.024	伐	0.022	逆	0.023	克	0.04	与	0.04	至
0.027	生	0.019	食	0.015	闻	0.023	许	0.016	用	0.021	围	0.038	命	0.04	闻
0.024	失	0.014	灭	0.014	辞	0.021	御	0.016	葬	0.019	待	0.028	请	0.04	入
0.023	谋	0.011	举	0.013	共	0.016	帅	0.016	说	0.018	还	0.023	听	0.033	出
0.023	卜	0.01	平	0.012	反	0.015	侵	0.015	乘	0.018	逐	0.023	退	0.019	适
0.02	守	0.01	敬	0.01	好	0.014	惧	0.015	聘	0.018	叛	0.022	往	0.018	弃
0.019	攻	0.009	处	0.009	当	0.014	信	0.015	归	0.015	入	0.018	去	0.015	成
0.017	求	0.009	封	0.007	毁	0.014	讨	0.014	拜	0.014	召	0.017	欲	0.012	致
0.016	赋	0.009	保	0.007	遇	0.013	能	0.013	免	0.014	受	0.017	辞	0.011	废
0.015	与	0.009	服	0.007	征	0.012	害	0.013	过	0.014	来	0.015	图	0.01	逃

最终,本文构建了一个含有军事行动、人口流动、社会动乱、政治外交、结盟议和、死亡丧葬、婚姻生育、政权更替、社会交换、生活风俗等十个大类,共 26 个小类的事件触发词分类体系,具体的类目及分析体系结构整理见表 3。

4.3 触发动词种子词集获取结果

对种子词集的构建,一方面是将动词聚类实验的结果中与主题具有强相关性的事件触发动词集合;另一方面将不同主题下同义词统计整理并将其整合到种子词集,完成种子词集的构建过程,构建样例见表 3。本文共采用三种实验方法进行词集扩展对比实验。

表 3 《左传》事件触发词种子词集示例

大类	小类	种子词示例	大类	小类	种子词示例
A 军事行动	A1 战争	伐、侵、御、帅	G 婚姻生育	G1 生育	生
	A2 阅兵	缮、阅、简		G2 婚姻	嫁、娶、逆
	A3 会师	会、遇		G3 通奸	通、嬖
B 人口流动	B1 逃跑	出、奔、逐	H 政权更替	H1 册立	立、封
	B2 归国	归、还、纳		H2 即位	即位、摄
C 社会动乱	C1 杀戮	杀、弑	I 社会交换	I1 贿赂	赂、献
	C2 叛乱	叛、乱		I2 赠与	赠、馈
D 政治外交	D1 朝见	见、召		I3 借取	假、借
	D2 聘问	聘、来聘	J 生活风俗	J1 祭祀	祀、祭
E 结盟议和	E1 盟会	会、盟		J2 占卜	卜
	E2 议和	平、求成		J3 狩猎	狩、田
F 死亡丧葬	F1 去世	卒、薨、缢		J4 建筑	城
	F2 丧仪	丧、葬		J5 疾病	病、疾

(1)第一次实验对没有标注词性的原文短句进行比较,计算相似度,输出每一句所对应的相似度前 10 的句子,生成了一个 17140 * n 的矩阵,计算结果见表 4。

(2)第二次实验在第一次实验的基础上加入了词性特征,计算结果见表 5。

(3)第三次实验对原文中抽取的 2305 个动词单独比较,计算相似度生成矩阵,进行相似度计算,计算结果见表 6。

通过三次对比实验得出,综合动词上下文特征和词本身的属性特征方案可以提高近义词识别的准确度,提升词集扩展的效果。

4.4 一致性检验

根据触发词的识别结果对《左传》的事件句进行了分类标注和结构化表示,运用 Kappa 系数对标注的结果进行了一致性计算,证明了触发词数据集的可信度。

4.4.1 事件句分类标注

事件一般由事件、地点、参与者三要素构成,涉及到的句子要素主要包括:时间状语、地点状语、主语、谓语、宾语。对被赋予类别的事件句进行结构化的表示,表 7 是事件句结构化标注示例:

4.4.2 一致性计算

本文选取 Kappa 系数作为一致性检验的计量指标。kappa 系数的计算结果在 0 ~ 0.2 之间时说明结果一致性极低(slight)、在 0.21 ~ 0.4 之间时说明结果的

表 4　未标注词性时的相似度计算结果表

相似度值句子序号	1	2	3	4	5	6	7	8	9	10
1	1.000	0.900	0.925	0.927	0.949	0.901	0.930	0.887	0.910	0.895
2	0.900	1.000	0.882	0.918	0.925	0.906	0.906	0.899	0.937	0.902
3	0.925	0.882	1.000	0.917	0.921	0.908	0.935	0.874	0.884	0.889
4	0.927	0.918	0.917	1.000	0.962	0.911	0.937	0.894	0.936	0.908
5	0.949	0.925	0.921	0.962	1.000	0.926	0.935	0.914	0.953	0.921
6	0.901	0.906	0.908	0.911	0.926	1.000	0.941	0.937	0.897	0.950
7	0.930	0.906	0.935	0.937	0.935	0.941	1.000	0.941	0.917	0.943
8	0.887	0.899	0.874	0.894	0.914	0.937	0.941	1.000	0.904	0.952
9	0.910	0.937	0.884	0.936	0.953	0.897	0.917	0.904	1.000	0.921
10	0.895	0.902	0.889	0.908	0.921	0.950	0.943	0.952	0.921	1.000

表 5　标注词性时的相似度计算结果表

相似度值句子序号	1	2	3	4	5	6	7	8	9	10
1	1.000	0.954	0.898	0.934	0.972	0.875	0.920	0.871	0.916	0.857
2	0.954	1.000	0.916	0.951	0.967	0.900	0.939	0.895	0.914	0.873
3	0.898	0.916	1.000	0.929	0.910	0.914	0.939	0.896	0.862	0.887
4	0.934	0.951	0.929	1.000	0.952	0.899	0.941	0.886	0.897	0.884
5	0.972	0.967	0.910	0.952	1.000	0.901	0.935	0.887	0.948	0.876
6	0.875	0.900	0.914	0.899	0.901	1.000	0.910	0.928	0.892	0.924
7	0.920	0.939	0.939	0.941	0.935	0.910	1.000	0.907	0.881	0.896
8	0.871	0.895	0.896	0.886	0.887	0.928	0.907	1.000	0.875	0.943
9	0.916	0.914	0.862	0.897	0.948	0.892	0.881	0.875	1.000	0.869
10	0.857	0.873	0.887	0.884	0.876	0.924	0.896	0.943	0.869	1.000

表 6　动词的相似度计算结果表

相似度值句子序号	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.000	0.850	0.915	0.915	0.915	0.889	0.888	0.837	0.910	0.915	0.785	0.903	0.904
2	0.850	1.000	0.841	0.841	0.841	0.824	0.834	0.781	0.822	0.841	0.718	0.805	0.837
3	0.915	0.841	1.000	1.000	1.000	0.897	0.910	0.854	0.908	1.000	0.784	0.910	0.914
4	0.915	0.841	1.000	1.000	1.000	0.897	0.910	0.854	0.908	1.000	0.784	0.910	0.914
5	0.915	0.841	1.000	1.000	1.000	0.897	0.910	0.854	0.908	1.000	0.784	0.910	0.914
6	0.889	0.824	0.897	0.897	0.897	1.000	0.933	0.918	0.929	0.897	0.792	0.931	0.929
7	0.888	0.834	0.910	0.910	0.910	0.933	1.000	0.893	0.932	0.910	0.796	0.928	0.946
8	0.837	0.781	0.854	0.854	0.854	0.918	0.893	1.000	0.878	0.854	0.739	0.893	0.890
9	0.910	0.822	0.908	0.908	0.908	0.929	0.932	0.878	1.000	0.908	0.811	0.955	0.967
10	0.915	0.841	1.000	1.000	1.000	0.897	0.910	0.854	0.908	1.000	0.784	0.910	0.914
11	0.785	0.718	0.784	0.784	0.784	0.792	0.796	0.739	0.811	0.784	1.000	0.800	0.817
12	0.903	0.805	0.910	0.910	0.910	0.931	0.928	0.893	0.955	0.910	0.800	1.000	0.965
13	0.904	0.837	0.914	0.914	0.914	0.929	0.946	0.890	0.967	0.914	0.817	0.965	1.000

表 7　事件句结构化表示示例

例句	八月，纪人伐夷。
时间	八月
地点	/
主语	纪人
谓语	伐
宾语	夷

一致性一般(fair)、在 0.41～0.6 之间时认为结果一致性中等(moderate)、在 0.61～0.8 之间时认为结果具有高度一致性(substantial)，当其趋近于 1 时，认为结果接近完全一致(almost perfect)。Kappa 系数的计算公式如下：

$$k = \frac{p - p}{1 - p}$$

$$p = \frac{a \times b + a \times b + \cdots a \times b}{n \times n}$$

其中,p0 是每一类正确分类的样本数量之和除以总样本数,也就是总体分类精度;n 为总样本个数;每一类的真实样本个数分别为 a1,a2,…,ax,而预测出来的每一类的样本个数分别为 b1,b2,…,bx。

邀请三位标引人员对 17140 条文本标注,通过对标引结果进行统计后计算其 Kappa 系数,计算结果为 0.74,处于 0.61 - 0.8 之间,说明本文所构建的事件触发动词集与事件句语义数据集是有效可信的。

6 总结及应用展望

本文基于多种文本挖掘技术探索建立面向典籍事件抽取的触发动词数据集的方法和技术。从构建方法上,本文运用 LDA 聚类模型进行主题聚类,结合定性分析方法,构建了面向典籍的触发动词的分类体系。从构建结果上,本文在细粒度的字词知识单元层面,建立了小规模的典籍事件触发词数据集,基于词典资源和基于机器学习两种方式对种子词集进行了扩展,对触发词分类体系进行了内容的丰富与填充。初步构建的数据集对于典籍事件抽取与识别提供了标注训练集。基于该数据集,研究采用 Bi-LSTM 方法[31]及 RoBERTa-CRF 方法[32]对同类型文本进行了事件抽取实验,数据集为上下文特征获取提供了统计参考依据,取得了较好的实验效果。本文的工作为今后开展大规模典籍内容挖掘提供了方法论参考。

本文在分类体系构建及数据集扩充的研究方法上还存在一些不足之处。首先,典籍事件的元素和实体众多,且不同事件涉及的同一实体、同一事件的不同实体之间往往存在着密切联系,由于研究的时间有限,本研究并没有对它们之间的联系做进一步的探究和阐释,在后续研究中将进一步进行探索。其次,对于语义数据集的有效性和可信度的评估时,仅仅通过 Kappa 系数这一指标来说明,后续研究中可以探索更多的评估指标,建立一个全面的数据集及分类体系评估系统,例如通过机器学习算法中的多元分类模型对标注数据进行试验比对。最后,受限于精加工古文本语料库的缺乏,本文的实验数据偏少,在后续的研究中将扩大处理样本的规模和数量,对上古、中古和近古不同历史时期的文本进行对比,扩大构建的触发动词数据集的数量和覆盖度,为典籍文本内容深度加工提供数据支撑。

参考文献:

[1] 黄水清,王东波. 古文信息处理研究的现状及趋势[J]. 图书情报工作,2017,61(12):43 - 49.

[2] SAMPO PYYSALO, TOMOKO OHTA, MAKOTO MIWA, et al. Event extraction across multiple levels of biological organization [J]. Bioinformatics, 2012, 28(18):i575 - i581.

[3] 黄佳艳. 面向金融新闻文本的事件识别与抽取[D]. 南京:东南大学,2019.

[4] 丁效,宋凡,秦兵,刘挺. 音乐领域典型事件抽取方法研究[J]. 中文信息学报,2011,25(2):15 - 20.

[5] 张海涛,李佳玮,刘伟利,等. 重大突发事件事理图谱构建研究[J]. 图书情报工作,2021,65(18):133 - 140.

[6] BUYKO E, FACSSLCR E, WCRMTCRJ, et al. Event extraction from trimmed dependency graphs[C]//Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Oregon: Association for Computational Linguistics, 2009: 19 - 27.

[7] VLACHOS A, BUTTERY P, SCAGHDHA D O, et al. Biomedical event extraction without training data[C]//Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Oregon: Association for Computational Linguistics, 2009: 7 - 10.

[8] 付剑锋. 面向事件的知识处理研究[D]. 上海:上海大学, 2010.

[9] MINH Q L, TRUONG S N, BAO Q H. A pattern approach for biomedical event annotation[C]//Proceedings of the BioNLP shared task 2011 workshop. Oregon: Association for Computational Linguistics, 2011:199 - 150.

[10] 张建海. 基于深度学习的生物医学事件抽取研究[D]. 大连:大连理工大学,2016.

[11] COHCN K B, VCRSPOOR K, JOHNSON H L, et al. High-precision biological event extraction with a concept recognizer[C]//Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Oregon: Association for Computational Linguistics, 2009: 50 - 58.

[12] BJORNE J, HEIMONEN J, UINTCR F, et al. Extracting complex biological events with rich graph-based feature sets[C]//Proceedings of the workshop on current trends in biomedical natural language processing: shared task. Oregon: Association for Computational Linguistics,2009:10 - 18.

[13] 陈箫箫, 刘波. 微博中的开放域事件抽取[J]. 计算机应用与软件,2016,33(8):18 - 22,109.

[14] 景悦诚, 黄征. 基于语言特征的舆情事件抽取[J]. 信息安全与通信保密,2015, 256(4):96 - 100.

[15] VLACHOS A, CRAVEN M. Biomedical event extraction from abstracts and full of papers using search based structured prediction [J]. BMC bio-informatics. 2012, 13 (Suppl 11): S5.

[16] 邓三鸿,胡昊天,王昊,等. 古文自动处理研究现状与新时代发展趋势展望[J]. 科技情报研究,2021,3(1):1 - 20.

[17] 邱冰,皇甫娟. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息,2008,24 (24):100 - 102.

[18] 徐润华, 陈小荷. 一种利用注疏的《左传》分词新方法[J]. 中文信息学报,2012,26(2):13 - 17,45.

[19] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京师范大学,2014.
[20] CHEN T, ZHU W, LV X, et al. A kalman filter based human-computer interactive word segmentation system for ancient Chinese texts [M]. Chinese computational linguistics and natural language processing based on naturally annotated big data. Berlin: Springer, 2013:25 -35.
[21] 黄建年. 农业古籍的计算机断句标点与分词标引研究[D]. 南京:南京农业大学,2009.
[22] 陈小荷, 冯敏萱, 徐润华, 等. 先秦文献信息处理[M]. 北京: 世界图书出版公司北京公司, 2013.
[23] 董志翘. 为中古汉语研究夯实基础——"中古汉语研究型语料库"建设琐议[J]. 燕山大学学报(哲学社会科学版), 2011, 12(01):1 -6.
[24] 王东波,高瑞卿,沈思,等. 面向先秦典籍的历史事件基本实体构件自动识别研究[J]. 国家图书馆学刊,2018,27(1):65 -77.
[25] 刘忠宝,党建飞, 张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J]. 图书情报工作,2020,64(11):116 -124.
[26] Linguistic Data Consortium. ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events [EB/OL]. [2021 -10 -16]. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/chinese-events-guidelines-v5.5.1.pdf.
[27] 纪国泰. 先秦汉语词汇研究的力作——评毛远明的《左传词汇研究》[J]. 成都师专学报,2000(1):74 -77.
[28] 孙丽丽. 春秋时期词汇研究[D]. 济南:山东大学,2012.
[29] SCHMIDT B M. Words alone: dismantling topic models in the humanities [J]. Journal of digital humanities, 2012, 2(1): 49 -65.
[30] UNDERWOOD T. What kinds of "topics" does topic modeling actually produce [EB/OL][2021 -03 -05]. http://tedunderwood.com/2012/04/01/what-kinds-oftopics-does-topic-modeling-actually-produce/.
[31] 马晓雯,何琳,刘建斌,等. 基于 Bi-LSTM 的古籍事件句触发词分类方法研究[J]. 农业图书情报学报,2021,33(9):27 -36.
[32] 喻雪寒,何琳,徐健. 基于 RoBERTa-CRF 的古文历史事件抽取方法研究[J]. 数据分析与知识发现,2021,5(7):26 -35.

作者贡献说明:

何琳:提出研究思路,设计研究过程,论文撰写;
马晓雯:负责算法设计、数据标注、论文撰写;
喻雪寒: 负责数据标注、程序修改;
艾毓茜: 负责数据标注;
李章超: 负责数据标注;
高丹: 负责数据标注。

Research on Recognition of Verbs Triggered by Events in Ancient Classics: Textual Experiments Based on *Zuo Zhuan*

He Lin[1,2] Ma Xiaowen[1,3] Yu Xuehan[1,2] Ai Yuxi[1,2] Li Zhangchao[1,2] Gao Dan[1,2]
[1] School of Information Management, Nanjing Agricultural University, Nanjing 210095
[2] Center for Humanities and Social Computational Lab of Nanjing Agricultural University, Nanjing 210095
[3] Nanjing Medical University Library, Nanjing 210029

Abstract: [**Purpose/significance**] Automatic event recognition and extraction is an important topic in current research on topic mining of ancient classics. Among them, the recognition of event trigger words is a basic work, which determined the quality of event extraction. This article aims to explore the general methods of automatic recognition and classification of event trigger words in ancient classics. [**Method/process**] Firstly, we explored the method of trigger verb classification construction by LDA topic clustering, which was carried out on the ancient classics combined with qualitative analysis. After the classification schema was confirmed, we building a preliminary seeds set of trigger words based on the clustering results. Then we expanded the trigger verb seeds set by the semantic similarity calculation on the ancient classics text resources. In the experiment, we took *Zuo Zhuan* as the experiment data sources, which is an important ancient classics in the Period of Chunqiu. The experiment tested the results of trigger verb classification construction and the expanding efficiency of trigger verb from the seeds set. [**Result/conclusion**] The results show that the method proposed in this paper is feasible and effective, and the event trigger word set constructed based on this has a high degree of credibility. The sample size and scope of the experiment can be further expanded in the future.

Keywords: trigger word recognition topic clustering word set expansion classification system construction ancient classic text